

Chromatic Clustering in High Dimensional Space

Hu Ding Jinhui Xu

Department of Computer Science and Engineering
State University of New York at Buffalo
{huding, jinhui}@buffalo.edu

Abstract. In this paper, we study a new type of clustering problem, called *Chromatic Clustering*, in high dimensional space. Chromatic clustering seeks to partition a set of colored points into groups (or clusters) so that no group contains points with the same color and a certain objective function is optimized. In this paper, we consider two variants of the problem, chromatic k -means clustering (denoted as k -CMeans) and chromatic k -medians clustering (denoted as k -CMedians), and investigate their hardness and approximation solutions. For k -CMeans, we show that the additional coloring constraint destroys several key properties (such as the locality property) used in existing k -means techniques (for ordinary points), and significantly complicates the problem. There is no FPTAS for the chromatic clustering problem, even if $k = 2$. To overcome the additional difficulty, we develop a standalone result, called *Simplex Lemma*, which enables us to efficiently approximate the mean point of an unknown point set through a fixed dimensional simplex. A nice feature of the simplex is its independence with the dimensionality of the original space, and thus can be used for problems in very high dimensional space. With the simplex lemma, together with several random sampling techniques, we show that a $(1 + \epsilon)$ -approximation of k -CMeans can be achieved in near linear time through a sphere peeling algorithm. For k -CMedians, we show that a similar sphere peeling algorithm exists for achieving constant approximation solutions.

1 Introduction

Clustering is one of the most fundamental problems in computer science and finds applications in many different areas [2–4, 6, 7, 9–12, 14, 16]. Most existing clustering techniques assume that the to-be-clustered data items are independent from each other. Thus each data item can “freely” determine its membership within the resulting clusters, without paying attention to the clustering of other data items. In recent years, there are also considerable attentions on clustering dependent data and a number of clustering techniques, such as correlation clustering, point-set clustering, ensemble clustering, and correlation connected clustering, have been developed [4, 7, 9–11].

In this paper, we consider a new type of clustering problems, called *Chromatic Clustering*, for dependent data. Roughly speaking, a chromatic clustering problem takes as input a set of colored data items and groups them into clusters, according to certain objective functions, so that no pair of items with the same color are grouped together (such a requirement is called *chromatic constraint*). Chromatic clustering captures the mutual exclusiveness relationship among data items and is a rather useful model for various applications. Due to the additional chromatic constraint, chromatic clustering is thus expected to simultaneously solve the “coloring” and clustering problems, which significantly complicates the problem. As it will be shown later, the chromatic clustering problem is challenging to solve even for the case that each color is shared only by two data items.

For chromatic clustering, we consider in this paper two variants, *Chromatic k -means Clustering* (k -CMeans) and *Chromatic k -median Clustering* (k -CMedians), in \mathbb{R}^d space, where the dimensionality could be very high and k is a fixed number. In both variants, the input is a set \mathcal{G} of n point-sets G_1, \dots, G_n with each containing a maximum of k points in d -dimensional space, and the objective is to partition all points of \mathcal{G} into k different clusters so that the chromatic constraint is satisfied and the total squared distance (i.e., k -CMeans) or total distance (i.e., k -CMedians) from each point to the center point (i.e., median or mean point) of its cluster is minimized.

Motivation: The chromatic clustering problem is motivated by several interesting applications. One of them is for determining the topological structure of chromosomes in cell biology [10]. In such applications, a set of 3D probing points (e.g., using BAC probes) is extracted from each homolog of the interested chromosome (see Figure 6 in Appendix), and the objective is to determine, for each chromosome homolog, the common spatial distribution pattern of the probes among a population of cells. For this purpose, the set of probes from each homolog is converted into a high dimensional feature point in the feature space, where each dimension represents the distance between a particular pair of probes. Since each chromosome has two (or more as in cancer cells) homologs, each cell contributes k (i.e., two or more) feature points. Due to technical limitation, it is impossible to identify the same homolog from all cells. Thus, the k feature points from each cell form a point-set with the same color (meaning that they are undistinguishable). To solve the problem, one could chromatically cluster all point-sets into k clusters (after normalizing the cell size), with each corresponding to a homolog, and use the mean or median point of each cluster as its common pattern.

Related works: As its generalization, chromatic clustering is naturally related to the traditional clustering problem. Due to the additional chromatic constraint, chromatic clustering could behave quite differently from its counterpart. For example, the k -means algorithms in [6, 15] relies on the fact that all input points in a Voronoi cell of the optimal k mean points belong to the same cluster. However, such a key locality property no longer holds for the k -CMeans problem.

Chromatic clustering falls in the umbrella of clustering with constraint. For such type of clustering, several solutions exist for some variants [5]. Unfortunately, due to their heuristic nature, none of them can yield quality guaranteed solutions for the chromatic clustering problem. The first quality guaranteed solution for chromatic clustering was obtained recently by Ding and Xu. In [10], they considered a special chromatic clustering problem, where every point-set has exactly k points in the first quadrant, and the objective is to cluster points by cones apexed at the origin, and presented the first PTAS for constant k . The k -CMeans and k -CMedians problems considered in this paper

are the general cases of the chromatic clustering problem. Very recently, Arkin *et al.* [1] considered a chromatic 2D 2-center clustering problem and presented both approximation and exact solutions.

1.1 Main Results and Techniques

In this paper, we present three main results, a constant approximation and a $(1 + \epsilon)$ -approximation for k -CMeans and their extensions to k -CMedians.

- **Constant approximation:** We show that given any c -approximation for k -means clustering, it could yield a $(2ck^2 + 2k - 1)$ -approximation for k -CMeans. This not only provides a way for us to generate an initial constant approximation solution for k -CMeans through some k -means algorithm, but more importantly reveals the intrinsic connection between the two clustering problems.
- **$(1 + \epsilon)$ -approximation:** We show that a near linear time $(1 + \epsilon)$ -approximation solution for k -CMeans can be obtained using an interesting sphere peeling algorithm. Due to the lack of locality property in k -CMeans, our sphere peeling algorithm is quite different from the ones used in [6, 15], which in general do not guarantee a $(1 + \epsilon)$ -approximation solution for k -CMeans as shown by our first result. Our sphere peeling algorithm is based on another standalone result, called *Simplex Lemma*. The simplex lemma enables us to obtain an approximate mean point of a set of unknown points through a grid inside a simplex determined by some partial knowledge of the unknown point set. A unique feature of the simplex lemma is that the complexity of the grid is *independent of the dimensionality*, and thus can be used to solve problems in high dimensional space. With the simplex lemma, our sphere peeling algorithm iteratively generates the mean points of k -CMeans with each iteration building a simplex for the mean point.
- **Extensions to k -CMedians:** We further extend the idea for k -CMeans to k -CMedians. Particularly, we show that any c -approximation for k -medians can be used to yield a $((2 + \epsilon)ck^2 + (2 + \epsilon)k + 1)$ -approximation for k -CMedians, where the ϵ error comes from the difficulty of computing the optimal median point (i.e., Fermat Weber point). With this and a similar sphere peeling technique, we obtain a $(5 + \epsilon)$ -approximation for k -CMedians. Note that although $k \geq 2$ is a constant in this paper, a $(5 + \epsilon)$ -approximation is still much better than a $((2 + \epsilon)ck^2 + (2 + \epsilon)k + 1)$ -approximation.

Due to space limit, many details of our algorithms, proofs, and figures are put in Appendix.

2 Preliminaries

In this section, we introduce some definitions which will be used throughout the paper.

Definition 1 (Chromatic Partition). Let $\mathcal{G} = \{G_1, \dots, G_n\}$ be a set of n point-sets with each $G_i = \{p_1^i, \dots, p_{k_i}^i\}$ consisting of $k_i \leq k$ points in \mathbb{R}^d space. A chromatic partition of \mathcal{G} is a partition of the $\sum_{1 \leq i \leq n} k_i$ points into k sets, U_1, \dots, U_k , such that each U_i contains no more than one point from each G_j for $j = 1, 2, \dots, n$.

Definition 2 (Chromatic k -means Clustering (k -CMeans)). Let $\mathcal{G} = \{G_1, \dots, G_n\}$ be a set of n point-sets with each $G_i = \{p_1^i, \dots, p_{k_i}^i\}$ consisting of $k_i \leq k$ points in \mathbb{R}^d space. The chromatic k -means clustering (or k -CMeans) of \mathcal{G} is to find k points $\{m_1, \dots, m_k\}$ in \mathbb{R}^d space and a chromatic partition U_1, \dots, U_k of \mathcal{G} such that $\frac{1}{n} \sum_j \sum_{q \in U_j} \|q - m_j\|^2$ is minimized. The problem is called full k -CMeans if $k_1 = k_2 = \dots = k_n = k$.

For both k -CMedians and k -CMeans, a problem often encountered in our approach is “How to find the best cluster for each point in G_i if the k mean or median points $A = \{m_1, \dots, m_k\}$ are already known?” An easy way to solve this problem is to first build a complete bipartite graph $(G_i \cup A, E_i)$ with points in G_i and A as the two partites and then compute a minimum weight bipartite matching as the solution, where the edge weight is the Euclidean distance or squared distance of the two corresponding vertices. Clearly, this can be done in a total of $O(k^3 dn)$ time for all G_i ’s. (We call this procedure as **bipartite matching**.)

3 Hardness of k -CMeans

It is easy to see that k -means is a special case of k -CMeans (i.e., each G_i contains exactly one point). As shown by Dasgupta [8], k -means in high dimensional space is NP-hard even if $k = 2$. Thus, we immediately have the following theorem.

Theorem 1. *k -CMeans is NP-hard for $k \geq 2$ in high dimensional space.*

3.1 Is Full k -CMeans Easier?

It is interesting to know whether full k -CMeans is easier than general k -CMeans, since it is disjoint with k -means when $k \geq 2$. The following theorem gives a negative answer to this question.

Theorem 2. *Full k -CMeans is NP-hard and has no FPTAS for $k \geq 2$ in high dimensional space unless $P=NP$ (see Appendix for the proof).*

The above theorem indicates that the fullness of k -CMeans does not reduce the hardness of the problem. However, this does not necessarily mean that full k -CMeans is as difficult as general k -CMeans to achieve a $(1 + \epsilon)$ -approximation for fixed k . Below we show that a $(1 + \epsilon)$ -approximation can be relatively easily achieved for full k -CMeans through some random sampling technique.

First we introduce a key lemma from [13]. Let S be a set of n points in \mathbb{R}^d space, T be a randomly selected subset from S with t points, and $\bar{x}(S)$, $\bar{x}(T)$ be the mean points of S and T respectively.

Lemma 1 ([13]). *With probability $1 - \eta$, $\|\bar{x}(S) - \bar{x}(T)\|^2 < \frac{1}{\eta t} \text{Var}^0(S)$, where $\text{Var}^0(S) = (\sum_{s \in S} \|s - \bar{x}(S)\|^2)/n$.*

Lemma 2. *Let S be a set of elements, and S' be a subset of S such that $\frac{|S'|}{|S|} = \alpha$. If randomly select $\frac{t \ln \frac{t}{\eta}}{\ln(1+\alpha)} = O(\frac{t}{\alpha} \ln \frac{t}{\eta})$ elements from S , with probability at least $1 - \eta$, the sample contains at least t elements from S' .*

Proof. If we randomly select z elements from S , then it is easy to know that with probability $1 - (1 - \alpha)^z$, there is at least one element from the sample belonging to S' . If we want the probability $1 - (1 - \alpha)^z$ equal to $1 - \eta/t$, z has to be $\frac{\ln \frac{t}{\eta}}{\ln \frac{1}{1-\alpha}} = \frac{\ln \frac{t}{\eta}}{\ln(1+\alpha)} \leq \frac{\ln \frac{t}{\eta}}{\ln(1+\alpha)} = O(\frac{1}{\alpha} \ln \frac{t}{\eta})$ (by Taylor series and $\alpha < 1$, $\ln(1 + \alpha) = O(\alpha)$). Thus if we perform t rounds of random sampling with each round selecting $O(\frac{1}{\alpha} \ln \frac{t}{\eta})$ elements, we get at least t elements from S' with probability at least $(1 - \eta/t)^t \geq 1 - \eta$. \square

Lemma 1 tells us that if we want to find an approximate mean point within a distance of $\epsilon \text{Var}^0(S)$ to the mean point, we just need to take a random sample of size $O(1/\epsilon)$. Lemma 2 suggests that for any set S and its subset $S' \subset S$ of size $\alpha|S|$, we can have a random subset T of S' with size $O(1/\epsilon)$ by randomly sampling directly from S $O(\frac{1}{\epsilon\alpha} \ln \frac{1}{\epsilon})$ points, even if S' is an unknown subset of S . Combining the two lemmas, we can immediately compute an approximation solution for full k -CMeans in the following way. First, we note that in full k -CMeans, each optimal cluster contains exact n points from the total of kn points in \mathcal{G} . This means that each cluster has a fraction of $\frac{1}{k}$ points from \mathcal{G} . Then, we can obtain an approximate mean point for each optimal cluster by (1) randomly sampling $O(\frac{k}{\epsilon} \ln \frac{1}{\epsilon})$ points from \mathcal{G} , (2) enumerating all possible subsets of size $O(1/\epsilon)$ to find the set T which is a random sample of the unknown optimal cluster, and (3) computing the mean of T as the approximate mean point of the optimal cluster. Finally, we can generate the k chromatic clusters from the k approximate mean points by using the bipartite matching procedure (see Section 2).

Theorem 3. *With constant probability, a $(1 + \epsilon)$ -approximation of full k -CMeans can be obtained in $O(2^{\text{poly}(\frac{k}{\epsilon})} nd)$ time.*

With the above theorem, we only need to focus on the general k -CMeans problem in the remaining sections. Note that in the general case, some clusters may have a very small fraction (rather than $1/k$) of points, thus we can not use the above method to solve the general k -CMeans problem.

4 Constant Approximation from k -means

In this section, we show that a constant approximation solution for k -CMeans can be produced from an approximation solution of k -means. Below is the main theorem of this section.

Theorem 4. Let $\mathcal{G} = \{G_1, \dots, G_n\}$ be an instance of k -CMeans, and \mathcal{C} be the k mean points of a constant c -approximation solution of k -means on the points $\cup_{i=1}^n G_i$. Then $[\mathcal{C}]^k$ contains at least one k -tuple which could induce a $(2ck^2 + 2k - 1)$ -approximation of k -CMeans on \mathcal{G} , where $[\mathcal{C}]^k = \underbrace{\mathcal{C} \times \dots \times \mathcal{C}}_k$.

To prove Theorem 4, we first introduce two lemmas.

Lemma 3. Let P be a set of points in \mathbb{R}^d space, and m be the mean point of P . For any point $m' \in \mathbb{R}^d$, $\sum_{p \in P} \|p - m'\|^2 = \sum_{p \in P} \|p - m\|^2 + |P| \times \|m - m'\|^2$ (see Appendix for the proof).

Lemma 4. Let P be a set of points in \mathbb{R}^d space, and P_1 be its subset containing $\alpha|P|$ points for some $0 < \alpha \leq 1$. Let m and m_1 be the mean points of P and P_1 respectively. Then $\|m_1 - m\| \leq \sqrt{\frac{1-\alpha}{\alpha}} \delta$, where $\delta^2 = \frac{1}{|P|} \sum_{p \in P} \|p - m\|^2$.

Proof. Let $P_2 = P \setminus P_1$, and m_2 be its mean point. By Lemma 3 we first have the following two equalities.

$$\sum_{p \in P_1} \|p - m\|^2 = \sum_{p \in P_1} \|p - m_1\|^2 + |P_1| \times \|m_1 - m\|^2. \quad (1)$$

$$\sum_{p \in P_2} \|p - m\|^2 = \sum_{p \in P_2} \|p - m_2\|^2 + |P_2| \times \|m_2 - m\|^2. \quad (2)$$

Then by the definition of δ , we have $\delta^2 = \frac{1}{|P|} (\sum_{p \in P_1} \|p - m\|^2 + \sum_{p \in P_2} \|p - m\|^2)$. Let $L = \|m_1 - m_2\|$. By the definition of mean point, we have $m = \frac{1}{|P|} \sum_{p \in P} p = \frac{1}{|P|} (\sum_{p \in P_1} p + \sum_{p \in P_2} p) = \frac{1}{|P|} (|P_1|m_1 + |P_2|m_2)$. Thus the three points $\{m, m_1, m_2\}$ are collinear, and $\|m_1 - m\| = (1 - \alpha)L$ and $\|m_2 - m\| = \alpha L$. Combining (1) and (2), we have

$$\begin{aligned} \delta^2 &= \frac{1}{|P|} \left(\sum_{p \in P_1} \|p - m_1\|^2 + |P_1| \times \|m_1 - m\|^2 + \sum_{p \in P_2} \|p - m_2\|^2 + |P_2| \times \|m_2 - m\|^2 \right) \\ &\geq \frac{1}{|P|} (|P_1| \times \|m_1 - m\|^2 + |P_2| \times \|m_2 - m\|^2) = \alpha((1 - \alpha)L)^2 + (1 - \alpha)(\alpha L)^2 = \alpha(1 - \alpha)L^2. \end{aligned}$$

Thus, we have $L \leq \frac{\delta}{\sqrt{\alpha(1-\alpha)}}$, which means that $\|m_1 - m\| = (1 - \alpha)L \leq \sqrt{\frac{1-\alpha}{\alpha}} \delta$. \square

Proof (of Theorem 4). Let $\{c_1, \dots, c_k\}$ be the k mean points in \mathcal{C} , and $\{S_1, \dots, S_k\}$ be their corresponding clusters. Let $\{m_1, \dots, m_k\}$ be the k unknown optimal mean points of k -CMeans, and $\mathcal{OPT} = \{Opt_1, \dots, Opt_k\}$ be the corresponding k optimal chromatic clusters. Let $\Gamma_j^i = Opt_i \cap S_j$, and τ_j^i be its mean point for $1 \leq i, j \leq k$. Since $\cup_{j=1}^k \Gamma_j^i = Opt_i$, by pigeonhole principle we know

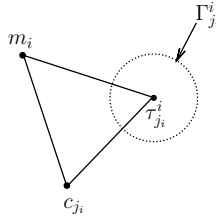


Fig. 1. An example illustrating Theorem 4.

that there must exist some index $1 \leq j_i \leq k$ such that $|\Gamma_{j_i}^i| \geq \frac{1}{k}|Opt_i|$. Thus by fixing j_i , we have the following about $\sum_{p \in Opt_i} \|p - c_{j_i}\|^2$ (see Figure 1)

$$\begin{aligned} \sum_{p \in Opt_i} \|p - c_{j_i}\|^2 &= \sum_{p \in Opt_i} \|p - m_i\|^2 + |Opt_i| \times \|m_i - c_{j_i}\|^2 = \sum_{p \in Opt_i} \|p - m_i\|^2 + |Opt_i| \times \|m_i - \tau_{j_i}^i + \tau_{j_i}^i - c_{j_i}\|^2 \\ &\leq \sum_{p \in Opt_i} \|p - m_i\|^2 + |Opt_i| \times (\|m_i - \tau_{j_i}^i\| + \|\tau_{j_i}^i - c_{j_i}\|)^2 \\ &\leq \sum_{p \in Opt_i} \|p - m_i\|^2 + |Opt_i| \times 2(\|m_i - \tau_{j_i}^i\|^2 + \|\tau_{j_i}^i - c_{j_i}\|^2), \end{aligned} \quad (3)$$

where the first equation follows from Lemma 3 (note that m_i is the mean point of Opt_i), and the last inequality follows from the fact that $(a+b)^2 \leq 2(a^2+b^2)$ for any numbers a and b . By Lemma 4, we have

$$\|\tau_{j_i}^i - m_i\|^2 \leq \frac{1 - \frac{1}{k}}{\frac{1}{k}} \left(\frac{1}{|Opt_i|} \sum_{p \in Opt_i} \|p - m_i\|^2 \right). \quad (4)$$

$$\|\tau_{j_i}^i - c_{j_i}\|^2 \leq \frac{1 - \frac{|\Gamma_{j_i}^i|}{|S_{j_i}|}}{\frac{|\Gamma_{j_i}^i|}{|S_{j_i}|}} \left(\frac{1}{|S_{j_i}|} \sum_{p \in S_{j_i}} \|p - c_{j_i}\|^2 \right). \quad (5)$$

Plugging (4) and (5) into inequality (3), we have

$$\begin{aligned} \sum_{p \in Opt_i} \|p - c_{j_i}\|^2 &\leq \sum_{p \in Opt_i} \|p - m_i\|^2 + |Opt_i| \times 2(\|m_i - \tau_{j_i}^i\|^2 + \|\tau_{j_i}^i - c_{j_i}\|^2) \\ &\leq \sum_{p \in Opt_i} \|p - m_i\|^2 + |Opt_i| \times 2 \left(\frac{1 - \frac{1}{k}}{\frac{1}{k}} \left(\frac{1}{|Opt_i|} \sum_{p \in Opt_i} \|p - m_i\|^2 \right) + \frac{1 - \frac{|\Gamma_{j_i}^i|}{|S_{j_i}|}}{\frac{|\Gamma_{j_i}^i|}{|S_{j_i}|}} \left(\frac{1}{|S_{j_i}|} \sum_{p \in S_{j_i}} \|p - c_{j_i}\|^2 \right) \right) \\ &= (2k-1) \sum_{p \in Opt_i} \|p - m_i\|^2 + 2 \frac{|Opt_i|}{|\Gamma_{j_i}^i|} \times \left(1 - \frac{|\Gamma_{j_i}^i|}{|S_{j_i}|} \right) \sum_{p \in S_{j_i}} \|p - c_{j_i}\|^2. \end{aligned}$$

Since $|\Gamma_{j_i}^i| \geq \frac{1}{k} |Opt_i|$, we have $\frac{|Opt_i|}{|\Gamma_{j_i}^i|} \times \left(1 - \frac{|\Gamma_{j_i}^i|}{|S_{j_i}|} \right) \leq k$. Thus the above inequality becomes

$$\sum_{p \in Opt_i} \|p - c_{j_i}\|^2 \leq (2k-1) \sum_{p \in Opt_i} \|p - m_i\|^2 + 2k \sum_{p \in S_{j_i}} \|p - c_{j_i}\|^2. \quad (6)$$

Summing both sides of (6) over i , we have

$$\begin{aligned} \sum_{i=1}^k \sum_{p \in Opt_i} \|p - c_{j_i}\|^2 &\leq (2k-1) \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\|^2 + 2k \sum_{i=1}^k \sum_{p \in S_{j_i}} \|p - c_{j_i}\|^2 \\ &\leq (2k-1) \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\|^2 + 2k^2 \sum_{j=1}^k \sum_{p \in S_j} \|p - c_j\|^2, \end{aligned} \quad (7)$$

where the second inequality follows from the inequality $\sum_{p \in S_{j_i}} \|p - c_{j_i}\|^2 \leq \sum_{j=1}^k \sum_{p \in S_j} \|p - c_j\|^2$, which implies that $2k \sum_{i=1}^k \sum_{p \in S_{j_i}} \|p - c_{j_i}\|^2 \leq 2k^2 \sum_{j=1}^k \sum_{p \in S_j} \|p - c_j\|^2$.

It is obvious that the optimal objective value of k -means is no larger than that of k -CMeans on the same set of points in \mathcal{G} . Thus, $\sum_{j=1}^k \sum_{p \in S_j} \|p - c_j\|^2 \leq c \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\|^2$. Plugging this inequality into inequality (7), we have

$$\sum_{i=1}^k \sum_{p \in Opt_i} \|p - c_{j_i}\|^2 \leq (2ck^2 + 2k - 1) \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\|^2.$$

The above inequality means that if we take the k -tuple $(c_{j_1}, \dots, c_{j_k})$ as the k approximate mean points for k -CMeans, we have a $(2ck^2 + 2k - 1)$ -approximation solution, where the k chromatic clusters can be obtained by the bipartite matching procedure. Thus, the theorem is proved. \square

Running Time: In the above theorem, the bipartite matching procedure takes $O(k^3 nd)$ time for one k -tuple. Since there are in total $O(k^k)$ such k -tuples, the total running time is $O(k^{k+3} nd)$ for computing a $(2ck^2 + 2k - 1)$ -approximation of k -CMeans from a c -approximation of k -means. As k is assumed to be a constant in this paper, the running time is linear.

5 $(1 + \epsilon)$ -Approximation Algorithm

This section presents our $(1 + \epsilon)$ -approximation solution to the k -CMeans problem. We first introduce a standalone result, *Simplex Lemma*, and then use it to achieve a $(1 + \epsilon)$ -approximation for k -CMeans. The main idea of the algorithm is to use a sphere peeling technique to generate the chromatic clusters iteratively, where the Simplex Lemma helps to determine a proper peeling region.

5.1 Simplex Lemma

Simplex Lemma is mainly for approximating the mean point of some **unknown** points set P . The only known information about P is a set S of j points with each of them being an approximate mean point of a subset of P . The following Simplex lemmas show that it is possible to construct a simplex of S and find the desired approximate mean point of P inside the simplex.

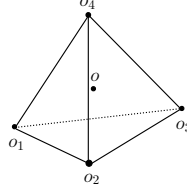


Fig. 2. An example for Lemma 5 with $j = 4$.

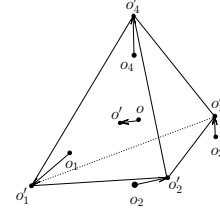


Fig. 3. An example for Lemma 6 with $j = 4$.

Lemma 5 (Simplex Lemma I). *Let P be a set of points in \mathbb{R}^d with a partition of $P = \cup_{l=1}^j P_l$ and $P_{l_1} \cap P_{l_2} = \emptyset$ for any $l_1 \neq l_2$. Let o be the mean point of P , and o_l be the mean point of P_l for $1 \leq l \leq j$. Further, let $\delta^2 = \frac{1}{|P|} \sum_{p \in P} \|p - o\|^2$, and V be the simplex determined by $\{o_1, \dots, o_j\}$. Then for any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V such that at least one grid point τ satisfies the inequality $\|\tau - o\| \leq \sqrt{\epsilon} \delta$.*

Proof. We will prove this lemma by mathematical induction on j .

Base case: For $j = 1$, since $P_1 = P$, $o_1 = o$. Thus, the simplex V and the grid are all simply the point o_1 . Clearly $\tau = o_1$ satisfies the inequality.

Induction step: Assume that the lemma holds for any $j \leq j_0$ for some $j_0 \geq 1$ (i.e., Induction Hypothesis). Now we consider the case of $j = j_0 + 1$. First, we assume that $\frac{|P_l|}{|P|} \geq \frac{\epsilon}{4j}$ for each $1 \leq l \leq j$. Otherwise, we can reduce the problem to the case of smaller j in the following way. Let $I = \{l | 1 \leq l \leq j, \frac{|P_l|}{|P|} < \frac{\epsilon}{4j}\}$ be the index set of small subsets. Then, $\frac{\sum_{l \in I} |P_l|}{|P|} < \frac{\epsilon}{4}$, and $\frac{\sum_{l \notin I} |P_l|}{|P|} \geq 1 - \frac{\epsilon}{4}$. By Lemma 4, we know that $\|o' - o\| \leq \sqrt{\frac{\epsilon/4}{1-\epsilon/4}} \delta$, where o' is the mean point of $\cup_{l \notin I} P_l$. Let $(\delta')^2$ be the variance of $\cup_{l \notin I} P_l$. Then, we have $(\delta')^2 \leq \frac{|P|}{|\cup_{l \notin I} P_l|} \delta^2 \leq \frac{1}{1-\epsilon/4} \delta^2$. Thus, if we replace P and ϵ by $\cup_{l \notin I} P_l$ and $\frac{\epsilon}{16}$ respectively, and find a point τ such that $\|\tau - o'\|^2 \leq \frac{\epsilon}{16} (\delta')^2 \leq \frac{\epsilon/16}{1-\epsilon/4} \delta^2$, we have $\|\tau - o\|^2 \leq (\|\tau - o'\| + \|o' - o\|)^2 \leq \frac{9}{16} \frac{\epsilon}{1-\epsilon/4} \delta^2 \leq \epsilon \delta^2$ (where the last inequality is due to the fact $\epsilon < 1$). This means that we can reduce the problem to a problem with point set $\cup_{l \notin I} P_l$ and a smaller j (i.e., $j - |I|$). By the induction hypothesis, we know that the reduced problem can be solved (note that the simplex would be a subset of V determined by $\{o_l | 1 \leq l \leq j, l \notin I\}$), and therefore the induction step holds for this case. Thus, in the following discussion, we can assume that $\frac{|P_l|}{|P|} \geq \frac{\epsilon}{4j}$ for each $1 \leq l \leq j$.

For each $1 \leq l \leq j$, since $\frac{|P_l|}{|P|} \geq \frac{\epsilon}{4j}$, by Lemma 4, we know that $\|o_l - o\| \leq \sqrt{\frac{1-\epsilon/4j}{\epsilon/4j}} \delta \leq 2\sqrt{\frac{j}{\epsilon}} \delta$. This, together with triangle inequality, implies that for any $1 \leq l, l' \leq j$, $\|o_l - o_{l'}\| \leq \|o_l - o\| + \|o_{l'} - o\| \leq 4\sqrt{\frac{j}{\epsilon}} \delta$. Thus, if we pick any index l_0 , and draw a ball \mathcal{B} centered at o_{l_0} and with radius $r = \max_{1 \leq l \leq j} \{\|o_l - o_{l_0}\|\} \leq 4\sqrt{\frac{j}{\epsilon}} \delta$, the whole simplex V will be inside \mathcal{B} . Note that since $o = \sum_{l=1}^j \frac{|P_l|}{|P|} o_l$, o also locates inside V . This indicates that we can construct \mathcal{B} in the $j-1$ -dimensional space spanned by $\{o_1, \dots, o_j\}$, rather than the whole \mathbb{R}^d space. Also, if we build a grid inside \mathcal{B} with grid length $\frac{\epsilon r}{4j}$, the total number of grid points is no more than $O((\frac{8j}{\epsilon})^j)$. With this grid, we know that for any point q inside V , there exists a grid point g such that $\|g - q\| \leq \sqrt{j(\frac{\epsilon r}{4j})^2} = \frac{\epsilon}{4\sqrt{j}} r \leq \sqrt{\epsilon} \delta$. This means that can find a grid point τ inside V , such that $\|\tau - o\|^2 \leq \epsilon \delta^2$. Thus, the induction step holds.

With the above base case and induction steps, the lemma holds for any $j \geq 1$. \square

In the above lemma, we assume that the exact positions of $\{o_1, \dots, o_j\}$ are known (see Fig. 2). However, in some scenario (e.g., the exact partition of P is not given, as is the case in k -CMeans), it is possible that we only know the approximate position of each mean point o_i (see Fig. 3). The following lemma shows that an approximate position of o can still be similarly determined.

Lemma 6 (Simplex Lemma II). *Let P , o , $P_l, o_l, 1 \leq l \leq j$, and, δ be defined as in Lemma 5. Let $\{o'_1, \dots, o'_j\}$ be j points in \mathbb{R}^d such that $\|o'_l - o_l\| \leq L$ for $1 \leq l \leq j$ and $L > 0$, and V' be the simplex determined by $\{o'_1, \dots, o'_j\}$. Then for any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V' such that at least one grid point τ satisfies the inequality $\|\tau - o\| \leq \sqrt{\epsilon}\delta + (1 + \epsilon)L$.*

5.2 Sphere Peeling Algorithm

This section presents a sphere peeling algorithm to achieve a $(1 + \epsilon)$ -approximation for k -CMeans.

Let $\mathcal{G} = \{G_1, \dots, G_n\}$ be an instance of k -CMeans with k (unknown) optimal chromatic clusters $\text{OPT} = \{\text{Opt}_1, \dots, \text{Opt}_k\}$, and m_j be the mean point of the cluster Opt_j for $1 \leq j \leq k$. Without loss of generality, we assume that $|\text{Opt}_1| \geq |\text{Opt}_2| \geq \dots \geq |\text{Opt}_k|$.

Algorithm overview: Our algorithm first computes a constant C -approximation solution (by Theorem 4) to determine an upper bound Δ of the optimal objective value δ_{opt}^2 , and then search for a good approximation of δ_{opt}^2 in the interval of $[\Delta/C, \Delta]$. At each search step, our algorithm performs a sphere peeling procedure to iteratively generate k approximate mean points for the chromatic clusters. Initially, the sphere peeling procedure uses random sampling technique (i.e., Lemma 1 and 2) to find an approximate mean point for Opt_1 . At $(j + 1)$ -th iteration, it already has approximate mean points $\{p_{v_1}, \dots, p_{v_j}\}$ for $\text{Opt}_1, \dots, \text{Opt}_j$ respectively. Then it draws j peeling spheres, $B_{j+1,1}, \dots, B_{j+1,j}$, centered at the j approximate mean points respectively and with a radius determined by the approximation of δ_{opt} . Denote the set of unknown points $\text{Opt}_{j+1} \setminus (\cup_{l=1}^j B_{j+1,l})$ as \mathcal{A} . Our algorithm considers two cases: (a) $|\mathcal{A}|$ is large enough and (b) $|\mathcal{A}|$ is small. For case (a), since $|\mathcal{A}|$ is large enough, we can first use Lemma 2 to find an approximate mean point $m_{\mathcal{A}}$ of \mathcal{A} , and then construct a simplex determined by $m_{\mathcal{A}}$ and $\{p_{v_1}, \dots, p_{v_j}\}$. For case (b), it directly constructs a simplex determined just by $\{p_{v_1}, \dots, p_{v_j}\}$. For either case, our algorithm builds a grid inside the simplex (i.e., using Lemma 6) to find an approximate mean point for Opt_{j+1} (i.e., $p_{v_{j+1}}$). Repeat the sphere peeling procedure k times to generate the k approximate mean points.

Algorithm k -CMeans

Input: $\mathcal{G} = \{G_1, \dots, G_n\}$, $k \geq 2$, and a small positive value ϵ .

Output: $(1 + \epsilon)$ -approximation solution for k -CMeans on \mathcal{G} .

1. Run the PTAS of k -means in [15] on \mathcal{G} , and let Δ be the obtained objective value.
2. For $i = 1$ to $\frac{2k}{\epsilon}$ do
 - (a) Set $\delta = \frac{\sqrt{\Delta}}{2k} + i \frac{\epsilon}{2k} \sqrt{\Delta}$, and run the Sphere-Peeling-Tree algorithm.
 - (b) Let \mathcal{T}_i be the output tree.
3. For each path of every \mathcal{T}_i , use bipartite matching procedure to compute the objective value of k -CMeans on \mathcal{G} . Output the k points from the path with the smallest objective value.

Algorithm Sphere-Peeling-Tree

Input: \mathcal{G} , $k \geq 2$, $\epsilon, \delta > 0$.

Output: A tree \mathcal{T} of height k with each node v associating with a point $p_v \in \mathbb{R}^d$.

1. Initialize \mathcal{T} with a single root node v associating with no point.
2. Recursively grow each node v in the following way
 - (a) If the height of v is already k , then it is a leaf.
 - (b) Otherwise, let j be the height of v . Build the radius candidates set $\mathcal{R} = \cup_{t=0}^{\log(kn)} \{ \frac{1+t\frac{\epsilon}{2}}{2(1+\epsilon)} j 2^{t/2} \sqrt{\epsilon} \delta \mid 0 \leq l \leq 4 + \frac{2}{\epsilon} \}$. For each $r \in \mathcal{R}$, do
 - i. Let $\{p_{v_1}, \dots, p_{v_j}\}$ be the j points associated with nodes on the root-to- v path.
 - ii. For each p_{v_l} , $1 \leq l \leq j$, construct a ball $B_{j+1,l}$ centered at p_{v_l} and with radius r .

- iii. Take a random sample from $\mathcal{G} \setminus \cup_{l=1}^j B_{j+1,l}$ with size $m = \frac{8k^3}{\epsilon^9} \ln \frac{k^2}{\epsilon^6}$. Compute the mean points of all subset of the sample, and denote them as $\Pi = \{\pi_1, \dots, \pi_{2^m-1}\}$.
- iv. For each $\pi_i \in \Pi$, construct the simplex determined by $\{p_{v_1}, \dots, p_{v_j}, \pi_i\}$. Also construct the simplex determined by $\{p_{v_1}, \dots, p_{v_j}\}$. Build a grid inside each simplex with size $O((\frac{32j}{\epsilon^2})^j)$.
- v. In total, there are $2^m (\frac{32j}{\epsilon^2})^j$ grid points inside the 2^m simplices. For each grid point, add one child to v , and associate it with the grid point.

Theorem 5. *With constant probability, Algorithm k -CMeans yields a $(1 + \epsilon)$ -approximation for k -CMeans in $O(2^{\text{poly}(\frac{k}{\epsilon})} n (\log n)^{k+1} d)$ time.*

5.3 Proof of Theorem 5

Let $\beta_j = |Opt_j| / |\cup_{i=1}^n G_i|$, and $\delta_j^2 = \frac{1}{|Opt_j|} \sum_{p \in Opt_j} \|p - m_j\|^2$, where m_j is the mean point of Opt_j . Clearly, $\beta_1 \geq \dots \geq \beta_k$ (by assumption) and $\sum_{j=1}^k \beta_j = 1$. Let $\delta_{opt}^2 = \sum_{j=1}^k \beta_j \delta_j^2$.

We prove Theorem 5 by mathematical induction. Instead of directly proving it, we consider the following two lemmas which jointly ensure the correctness of Theorem 5.

Lemma 7. *Among all the trees generated in Algorithm k -CMeans, with constant probability, there exists at least one tree, \mathcal{T}_i , which has a root-to-leaf path with each node v_j at level j , $1 \leq j \leq k$, on the path associating a point p_{v_j} and satisfying the inequality $\|p_{v_j} - m_j\| \leq \epsilon \delta_j + (1 + \epsilon) j \sqrt{\frac{\epsilon}{\beta_j}} \delta_{opt}$.*

Before proving this lemma, we first show its implication.

Lemma 8. *If Lemma 7 is true, Algorithm k -CMeans yields a $(1 + O(k^3)\epsilon)$ -approximation for k -CMeans.*

Proof. We first assume that Lemma 7 is true. Then for each $1 \leq j \leq k$, we have

$$\begin{aligned} \sum_{p \in Opt_j} \|p - p_{v_j}\|^2 &= \sum_{p \in Opt_j} \|p - m_j\|^2 + |Opt_j| \times \|m_j - p_{v_j}\|^2 \leq \sum_{p \in Opt_j} \|p - m_j\|^2 + |Opt_j| \times 2(\epsilon^2 \delta_j^2 + (1 + \epsilon)^2 j^2 \frac{\epsilon}{\beta_j} \delta_{opt}^2) \\ &= (1 + 2\epsilon^2) |Opt_j| \delta_j^2 + 2(1 + \epsilon)^2 j^2 \epsilon |\mathcal{G}| \delta_{opt}^2, \end{aligned} \quad (8)$$

where the first equation follows from Lemma 3 (note that m_j is the mean point of Opt_j), the second inequality follows from Lemma 7 and the fact that $(a + b)^2 \leq 2(a^2 + b^2)$ for any two real numbers a and b , and the last equality follows from $\frac{|Opt_j|}{\beta_j} = |\mathcal{G}|$. Summing both sides of (8) over j , we have

$$\begin{aligned} \sum_{j=1}^k \sum_{p \in Opt_j} \|p - p_{v_j}\|^2 &\leq \sum_{j=1}^k ((1 + 2\epsilon^2) |Opt_j| \delta_j^2 + 2(1 + \epsilon)^2 j^2 \epsilon |\mathcal{G}| \delta_{opt}^2) \\ &\leq (1 + 2\epsilon^2) \sum_{j=1}^k |Opt_j| \delta_j^2 + 2(1 + \epsilon)^2 k^3 \epsilon |\mathcal{G}| \delta_{opt}^2 = (1 + O(k^3)\epsilon) |\mathcal{G}| \delta_{opt}^2, \end{aligned} \quad (9)$$

where the last equation follows from the fact that $\sum_{j=1}^k |Opt_j| \delta_j^2 = |\mathcal{G}| \delta_{opt}^2$. By (9), we know that $\{p_{v_1}, \dots, p_{v_k}\}$ will induce a $(1 + O(k^3)\epsilon)$ -approximation solution for k -CMeans via bipartite matching procedure. Since Algorithm k -CMeans outputs the best solution generated in all trees, the resulting solution is clearly a $(1 + O(k^3)\epsilon)$ -approximation solution. Thus the lemma is true. \square

The above lemma indicates that if we replace ϵ by $\frac{\epsilon}{k^3}$ in the input of our algorithm, it will result in a $(1 + \epsilon)$ -approximation solution. This implies that Lemma 7 is indeed sufficient to ensure the correctness of Theorem 5 (except for the time complexity). Now we prove Lemma 7.

Proof (of Lemma 7). Note that $\Delta \leq 4k^2 \delta_{opt}^2$, and we build ϵ -net in $[\frac{\sqrt{\Delta}}{2k}, \sqrt{\Delta}]$. Let \mathcal{T}_i be the tree generated by Algorithm Sphere-Peeling-Tree and corresponding to the input $\delta \in [\delta_{opt}, (1 + \epsilon)\delta_{opt}]$. We will focus our discussion on \mathcal{T}_i , and prove the lemma by mathematical induction on j .

Base case: For $j = 1$, since $\beta_1 = \max\{\beta_j | 1 \leq j \leq k\}$, we have $\beta_1 \geq \frac{1}{k}$. By Lemmas 1 and 2, we can find the approximation mean point through random sampling. Let p_{v_1} be the approximation mean point. Clearly, $\|p_{v_1} - m_1\| \leq \epsilon \delta_1 \leq \epsilon \delta_1 + (1 + \epsilon) \sqrt{\frac{\epsilon}{\beta_1}} \delta_{opt}$ (By Lemmas 1 and 2).

Induction step: We assume that there is a path in \mathcal{T}_i from the root to the j_0 -th level, such that for each $1 \leq l \leq j_0$, the level- l node v_l on the path is associated with a point p_{v_l} satisfying the inequality $\|p_{v_l} - m_l\| \leq \epsilon \delta_l + (1 + \epsilon)l \sqrt{\frac{\epsilon}{\beta_l}} \delta_{opt}$ (i.e., Induction Hypothesis). Now we consider the case of $j = j_0 + 1$. Below we will show that there is one child of v_{j-1} , i.e., v_j , such that its associated point p_{v_j} satisfies the inequality $\|p_{v_j} - m_j\| \leq \epsilon \delta_j + (1 + \epsilon)j \sqrt{\frac{\epsilon}{\beta_j}} \delta_{opt}$. First, we have the following claim (see Appendix for the proof).

Claim (1). In the set of radius candidates built in Algorithm Sphere-Peeling-Tree, there exists one value $r_j \in \mathcal{R}$ such that

$$j \sqrt{\frac{\epsilon}{\beta_j}} \delta_{opt} \leq r_j \leq (1 + \frac{\epsilon}{2})j \sqrt{\frac{\epsilon}{\beta_j}} \delta_{opt}.$$

Now, we construct the $j - 1$ peeling spheres, $\{B_{j,1}, \dots, B_{j,j-1}\}$ (as in Algorithm Sphere-Peeling-Tree). For each $1 \leq l \leq j - 1$, $B_{j,l}$ is centered at p_{v_l} and with radius r_j . By Markov inequality and induction hypothesis, we have the following claim (see Appendix for the proof).

Claim (2). For each $1 \leq l \leq j - 1$, we have $|Opt_l \setminus (\bigcup_{w=1}^{j-1} B_{j,w})| \leq \frac{4\beta_j |\mathcal{G}|}{\epsilon}$.

Claim 2 shows that $|Opt_l \setminus (\bigcup_{w=1}^{j-1} B_{j,w})|$ is bounded for $1 \leq l \leq j - 1$, which helps us to find the approximate mean point of Opt_j . Induced by the $j - 1$ peeling spheres $\{B_{j,1}, \dots, B_{j,j-1}\}$, Opt_j is divided into j subsets, $Opt_j \cap B_{j,1}, \dots, Opt_j \cap B_{j,j-1}$ and $Opt_j \setminus (\bigcup_{w=1}^{j-1} B_{j,w})$. To simplify our discussion, we let P_l denote $Opt_j \cap B_{j,l}$ for $1 \leq l \leq j - 1$, P_j denote $Opt_j \setminus (\bigcup_{w=1}^{j-1} B_{j,w})$, and τ_l denote the mean point of P_l . Note that the peeling spheres may intersect with each other. For any two intersecting spheres B_{j,l_1} and B_{j,l_2} , we let the points set $Opt_j \cap (B_{j,l_1} \cap B_{j,l_2})$ belong to either P_{l_1} or P_{l_2} arbitrarily. Thus, we can assume that $\{P_l \mid 1 \leq l \leq j\}$ are pairwise disjoint. Now consider the size of P_j (i.e., $|P_j|$). We have the following two cases: (a) $|P_j| \geq \epsilon^3 \frac{\beta_j}{j} |\mathcal{G}|$ and (b) $|P_j| < \epsilon^3 \frac{\beta_j}{j} |\mathcal{G}|$. In the following, we show how, in each case, Algorithm Sphere-Peeling-Tree can obtain an approximate mean point for Opt_j by using the Simplex Lemma (i.e., Lemma 6).

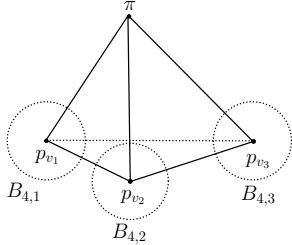


Fig. 4. Case (a) for $j = 4$.

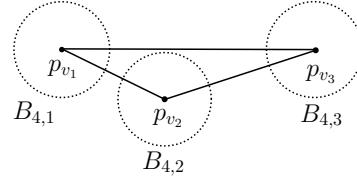


Fig. 5. Case (b) for $j = 4$.

For case (a), by Claim 2, together with the fact that $\beta_l \leq \beta_j$ for $l > j$, we know that

$$\frac{|P_j|}{\sum_{1 \leq i \leq k} |Opt_i \setminus (\bigcup_{l=1}^{j-1} B_{j,l})|} \geq \frac{\frac{\epsilon^3}{j} \beta_j}{\frac{4(j-1)\beta_j}{\epsilon} + \frac{\epsilon^2}{j} \beta_j + (k-j)\beta_j} > \frac{\epsilon^4}{8kj} \geq \frac{\epsilon^4}{8k^2}.$$

This means that P_j is large enough, comparing to the set of points outside the peeling spheres. Hence, we can use random sampling technique to obtain an approximate mean point π for P_j in the following way. First, we set $t = \frac{k}{\epsilon^5}$, $\eta = \frac{\epsilon}{k}$, and take a sample of size $\frac{t \ln(t/\eta)}{\epsilon^4/8k^2} = \frac{8k^3}{\epsilon^9} \ln \frac{k^2}{\epsilon^6}$. By Lemma 2, we know that with probability $1 - \frac{\epsilon}{k}$, the sample contains $\frac{k}{\epsilon^5}$ points from P_j . Then we let π be the mean point of the $\frac{k}{\epsilon^5}$ points from P_j , and a^2 be the variance of P_j . By Lemma 1, we know that with probability $1 - \frac{\epsilon}{k}$, $\|\pi - \tau_j\|^2 \leq \epsilon^4 a^2$. Also, since $\frac{|P_j|}{|Opt_j|} \geq \frac{\epsilon^3}{j}$, we have $a^2 \leq \frac{|Opt_j|}{|P_j|} \delta_j^2 \leq \frac{j}{\epsilon^3} \delta_j^2$. Thus, $\|\pi - \tau_j\|^2 \leq \epsilon j \delta_j^2$.

Once obtaining π , we can now use Lemma 6 to find a point p_{v_j} satisfying the condition of $\|p_{v_j} - m_j\| \leq \epsilon \delta_j + (1 + \epsilon)j \sqrt{\frac{\epsilon}{\beta_j}} \delta_{opt}$. First, we construct a simplex $V'_{(a)}$ determined by $\{p_{v_1}, \dots, p_{v_{j-1}}\}$ and π (see Figure. 4). Note that Opt_j is divided by the peeling spheres into j disjoint subsets, P_1, \dots, P_j , which is a partition of Opt_j . Each P_l ($1 \leq l \leq j - 1$) locates inside $B_{j,l}$, which implies that τ_l is also

inside $B_{j,l}$. Further, since $\|p_{v_l} - \tau_l\| \leq r_j \leq (1 + \frac{\epsilon}{2})j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}$ for $1 \leq l \leq j-1$ (by Claim 1), and $\|\pi - \tau_j\| \leq \sqrt{\epsilon j}\delta_j \leq \sqrt{\frac{\epsilon j}{\beta_j}}\delta_{opt}$ (by $\beta_j\delta_j^2 \leq \delta_{opt}^2$, which implies $\delta_j \leq \sqrt{1/\beta_j}\delta_{opt}$), after setting the value of L (in Lemma 6) to be $\max\{r_j, \|\pi - \tau_j\|\} \leq \max\{(1 + \frac{\epsilon}{2})j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}, \sqrt{\frac{\epsilon j}{\beta_j}}\delta_{opt}\} \leq (1 + \frac{\epsilon}{2})j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}$ and the value of ϵ (in Lemma 6) to be $\epsilon_0 = \epsilon^2/4$, by Lemma 6 we can construct a grid inside the simplex $V'_{(a)}$ with size $O((\frac{8j}{\epsilon_0})^j)$ which ensures the existence of one grid point τ satisfying the inequality of $\|\tau - m_j\| \leq \sqrt{\epsilon_0}\delta_j + (1 + \epsilon_0)L \leq \epsilon\delta_j + (1 + \epsilon)j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}$. Hence, we can use τ as p_{v_j} , and the induction step holds for this case.

For case (b), since P_j has a small size, we cannot directly perform random sampling on it to find its approximate mean point. To overcome this difficulty, we merge P_j with some other large subset P_l . Particularly, since $\sum_{l=1}^{j-1} |P_l| = |Opt_j| - |P_j| \geq (\beta_j - \epsilon^3 \frac{\beta_j}{j})|\mathcal{G}|$, by pigeonhole principle, we know that there exists one l_0 such that P_{l_0} has size at least $\frac{1}{j-1}(\beta_j - \epsilon^3 \frac{\beta_j}{j})|\mathcal{G}|$. Without loss of generality, we assume $l_0 = 1$. Then $|P_1| \geq \frac{1}{j-1}(\beta_j - \epsilon^3 \frac{\beta_j}{j})|\mathcal{G}|$, and we can view $P_1 \cup P_j$ as one large enough subset of Opt_j . Let τ' denote the mean point of $P_1 \cup P_j$, then we have the following claim (see Appendix for the proof).

Claim (3). $\|\tau_1 - \tau'\| \leq \frac{\sqrt{2}\epsilon}{1-\epsilon^3} \sqrt{\frac{j\epsilon}{\beta_j}}\delta_{opt}$.

This means that we can also use Lemma 6 to find an approximate mean point in a way similar to case (a) (see Figure. 5); the difference is that Opt_j is divided into $j-1$ subsets (i.e., P_1 and P_j is viewed as one subset $P_1 \cup P_j$) and the value of L is set to be $r_j + \|\tau_1 - \tau'\| \leq r_j + \frac{\sqrt{2}\epsilon}{1-\epsilon^3} \sqrt{\frac{j\epsilon}{\beta_j}}\delta_{opt}$. We can first construct a simplex $V'_{(b)}$ determined by $\{p_{v_1}, \dots, p_{v_{j-1}}\}$ (see Figure. 5), and then build a grid inside $V'_{(b)}$ with size $O((\frac{8j}{\epsilon_0})^j)$, where $\epsilon_0 = \epsilon^2/4$. By Lemma 6, we know that there exists one grid point τ satisfying the condition of $\|\tau - m_j\| \leq \sqrt{\epsilon_0}\delta_j + (1 + \epsilon_0)L \leq \epsilon\delta_j + (1 + \epsilon)j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}$. Thus the induction step holds for this case.

Since Algorithm Sphere-Peeling-Tree executes every step in our above discussion, the induction step, as well as the lemma, is true. \square

Success probability: From the above analysis, we know that in the j -th step/iteration, only case (a) (i.e., $|P_j| \geq \epsilon^3 \frac{\beta_j}{j} |\mathcal{G}|$) needs to consider success probability, since case (b) (i.e., $|P_j| < \epsilon^3 \frac{\beta_j}{j} |\mathcal{G}|$) does not need to do sampling. Recall that in case (a), we take a sample of size $\frac{8k^3}{\epsilon^9} \ln \frac{k^2}{\epsilon^6}$. Thus with probability $1 - \frac{\epsilon}{k}$, it contains $\frac{k}{\epsilon^5}$ points from P_j . Meanwhile, with probability $1 - \frac{\epsilon}{k}$, $\|\pi - \tau_j\|^2 \leq \epsilon^4 a^2$. Hence, the success probability in the j -th step is $(1 - \frac{\epsilon}{k})^2$, which means that the success probability in all k steps is $(1 - \frac{\epsilon}{k})^{2k} \geq 1 - 2\epsilon$.

Running time: Algorithm k -CMeans calls Algorithms Sphere-Peeling-Tree $\frac{2k}{\epsilon}$ times. It is easy to see that each node on the tree returned from Algorithm Sphere-Peeling-Tree has $|\mathcal{R}|2^m(\frac{32j}{\epsilon^2})^j$ children, where $|\mathcal{R}| = O(\frac{\log kn}{\epsilon})$, and $m = \frac{8k^3}{\epsilon^9} \ln \frac{k^2}{\epsilon^6}$. Since the tree has a height of k , the complexity of the tree is $O(2^{\text{poly}(\frac{k}{\epsilon})}(\log n)^k)$. Further, since each node takes $O(|\mathcal{R}|2^m(\frac{32j}{\epsilon^2})^j nd)$ time, the total time complexity of Algorithm k -CMeans is $O(2^{\text{poly}(\frac{k}{\epsilon})}n(\log n)^{k+1}d)$.

6 Extension to Chromatic k -Medians Clustering

We extend our ideas for k -CMeans to the Chromatic k -Medians Clustering problem (k -CMedians). Similar to k -CMeans, we first show its relationship with k -medians, and then present a $(5 + \epsilon)$ -approximation algorithm using the sphere peeling technique. Due to the lack of a similar Simplex Lemma for k -CMedians, we achieve a constant approximation, instead of a PTAS. See details of the algorithm in Section 14 of the Appendix.

References

1. Esther M. Arkin, Jos Miguel Daz-Bez, Ferran Hurtado, Piyush Kumar, Joseph S. B. Mitchell, Beln Palop, Pablo Prez-Lantero, Maria Saumell, Rodrigo I. Silveira: Bichromatic 2-Center of Pairs of Points. *LATIN 2012*: 25-36
2. P. K. Agarwal and C. M. Procopiuc. "Exact and Approximation Algorithms for Clustering," *Proc. 9th ACM-SIAM Sympos. Discrete Algorithms*, pages 658-667, 1998.
3. David Arthur, Sergei Vassilvitskii: "k-means++: the advantages of careful seeding". *SODA 2007*: 1027-1035
4. Nikhil Bansal, Avrim Blum, Shuchi Chawla: "Correlation Clustering". *Machine Learning* 56(1-3): 89-113 (2004)
5. S.Basu, Ian Davidson: Clustering with Constraints Theory and Practice. *ACM KDD 2006*
6. M.Badoiu, S.Har-Peled, P.Indyk, "Approximate clustering via core-sets", *Proceedings of the 34th Symposium on Theory of Computing*, pp. 250-257, 2002.
7. C.Blum, K.Kailing, P.Krger, A.Zimek, "Computing Clusters of Correlation Connected Objects". *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD'04)*, Paris, France. pp. 455-467. doi:10.1145/1007568.1007620.
8. S. Dasgupta, "The hardness of k-means clustering". *Technical Report*, 2008.
9. Erik Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. "Correlation clustering in general weighted graphs". *Theor. Comput. Sci.*, 361(2):172-187, 2006
10. H.Ding, J.Xu, "Solving Chromatic Cone Clustering via Minimum Spanning Sphere", *ICALP*, 2011
11. Ioannis Giotis and Venkatesan Guruswami. "Correlation clustering with a fixed number of clusters". *Theory of Computing*, 2(1):249-266, 2006.
12. S. Har-Peled and S. Mazumdar, "Coresets for k-Means and k-Median Clustering and their Applications," *Proc. 36th ACM Symposium on Theory of Computing*, pages 291-300, 2004.
13. Mary Inaba, Naoki Katoh, Hiroshi Imai, "Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering (Extended Abstract)". *Symposium on Computational Geometry 1994*: 332-339
14. S. G. Kolliopoulos and S. Rao, "A nearly linear-time approximation scheme for the euclidean k-median problem," *Proc. 7th Annu. European Sympos. Algorithms*, pages 378-389, 1999.
15. A. Kumar, Y. Sabharwal, S. Sen, "Linear-time approximation schemes for clustering problems in any dimensions". *J. ACM* 57(2):2010
16. R.Ostrovsky, Y.Rabani, L.J.Schulman, and C.Swamy. "The Effectiveness of Lloyd-Type Methods for the k-Means Problem". *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. pp. 165-174.

7 Figure. 6

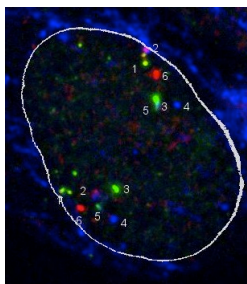


Fig. 6. BAC probes of Chromosome 1 in a WI38 cell with homolog having 6 probes.

8 Proof for Theorem 2

Proof. Since it is sufficient to show that the theorem holds for the case of $k = 2$, we assume in this proof that $k = 2$ and each point-set G_i has exactly two points. We make use of a construction by Dasgupta for the NP-hardness proof of the 2-mean clustering problem in high dimensional space [8]. Their proof reduces from the NAE3SAT problem. For better understanding our ideas, below we sketch their construction.

1. For any instance ϕ of NAE3SAT with literal set $\{x_1, \dots, x_n\}$ and m clauses, construct a $2n \times 2n$ matrix $D_{\alpha,\beta}$ as follows, where the indices correspond to $\{x_1, \dots, x_n\}$ when they are in the range

of $[1, n]$, and to $\{\overline{x_1}, \dots, \overline{x_n}\}$ when they are in the range of $[n+1, 2n]$.

$$D_{\alpha, \beta} = \begin{cases} 0 & \text{if } \alpha = \beta \\ 1 + \Delta & \text{if } \alpha = \overline{\beta} \\ 1 + \delta & \text{if } \alpha \sim \beta \\ 1 & \text{otherwise,} \end{cases} \quad \text{where } \Delta, \delta \text{ are two constants satisfying inequalities } 0 < \delta < \Delta < 1$$

and $4\delta m < \Delta \leq 1 - 2\delta n$, and $\alpha \sim \beta$ means that both α and β or both $\overline{\alpha}$ and $\overline{\beta}$ appear in a clause.

2. D can be embedded into R^{2n} , i.e., there exist $2n$ points in R^{2n} with D as their distance matrix.
3. Let C_1 and C_2 be the two clusters of the 2-mean clustering of the $2n$ embedding points. If for any i , the points corresponding to x_i and $\overline{x_i}$ are separated into different clusters, then ϕ is satisfiable if and only if

$$\frac{1}{2n} \sum_{i,j \in C_1} D_{i,j} + \frac{1}{2n} \sum_{i,j \in C_2} D_{i,j} \leq n - 1 + \frac{2\delta m}{n}.$$

4. Since $\frac{1}{2n} \sum_{i,j \in C_1} D_{i,j} + \frac{1}{2n} \sum_{i,j \in C_2} D_{i,j}$ is the total cost of the 2-mean clustering for C_1 and C_2 , a polynomial time solution to the 2-mean clustering problem in high dimensional space implies a polynomial time solution to NAE3SAT. Thus the 2-mean clustering is NP-hard in high dimensions.

The above reduction can be naturally extended to show the NP-hardness of the full chromatic 2-mean clustering problem. To show this, we only need to construct G_i as the set containing the two points corresponding to x_i and $\overline{x_i}$ (for simplicity, we write it as $G_i = \{x_i, \overline{x_i}\}$), and the remaining proof follows from the same argument.

Next, we show that full 2-CMean has no FPTAS in high dimensional space unless $P=NP$. To see this, we still use the same construction. From the above discussion, we know that ϕ is unsatisfiable if and only if for any chromatic partition of \mathcal{G} , there exists one clause in ϕ such that the three points corresponding to the three literals in this clause are clustered into the same cluster. Hence, the total cost for any chromatic partition is at least

$$2\frac{1}{n} \binom{n}{2} + (m-1)\delta + 3\delta = n - 1 + \frac{2}{n}(m+2)\delta.$$

The ratio η between the minimum chromatic partition cost of an unsatisfiable instance and the upper bound cost of a satisfiable instance is

$$\eta = \frac{n - 1 + \frac{2}{n}(m+2)\delta}{n - 1 + \frac{2\delta m}{n}} = 1 + \frac{\frac{4}{n}\delta}{n - 1 + \frac{2(m+2)}{n}\delta}.$$

If we let $\delta = \frac{1}{5m+2n}$, then $\eta = 1 + \frac{\frac{4}{n}\delta}{n - 1 + \frac{2(m+2)}{n}\delta} = 1 + \frac{4}{n(5m+2n)(n-1) + 2(m+2)}.$

Suppose that there exists an FPTAS for the full chromatic 2-means clustering problem. Then, if we let $\epsilon < \frac{4}{n(5m+2n)(n-1) + 2(m+2)}$, the cost of a $(1 + \epsilon)$ -approximation of the full 2-CMeans is less than $n - 1 + \frac{2}{n}(m+2)\delta$ if and only if ϕ is satisfiable. Since the running time of the FPTAS for full 2-CMeans and $\frac{1}{\epsilon}$ are all polynomial functions of m and n , this implies that NAE3SAT can be solved in polynomial time. Obviously this can only happen if $P=NP$. \square

9 Proof of Lemma 3

Proof. In the our following discussion, we use $\langle a, b \rangle$ to denote the inner product of a and b . It is easy to see that

$$\sum_{p \in P} \|p - m'\|^2 = \sum_{p \in P} \|p - m + m - m'\|^2$$

$$\begin{aligned}
&= \sum_{p \in P} (||p - m||^2 + 2 \langle p - m, m - m' \rangle + ||m - m'||^2) \\
&= \sum_{p \in P} ||p - m||^2 + 2 \sum_{p \in P} \langle p - m, m - m' \rangle + |P| \times ||m - m'||^2 \\
&= \sum_{p \in P} ||p - m||^2 + 2 \langle \sum_{p \in P} (p - m), m - m' \rangle + |P| \times ||m - m'||^2.
\end{aligned}$$

Since m is the mean point of P , $\sum_{p \in P} (p - m) = 0$. Thus, the above equality becomes $\sum_{p \in P} ||p - m'||^2 = \sum_{p \in P} ||p - m||^2 + |P| \times ||m - m'||^2$. \square

10 Proof of Lemma 6

Proof. Similar to Lemma 5, we prove this lemma by mathematics induction on j .

Base case. For $j = 1$, since $o_1 = o$, we just need to let $\tau = o'_1$. Then, we have $||\tau - o|| = ||o'_1 - o|| = ||o'_1 - o_1|| \leq L \leq \sqrt{\epsilon}\delta + (1 + \epsilon)L$. Thus, the base case holds.

Induction step. Assume that the lemma holds for any $j \leq j_0$ for some $j_0 \geq 1$ (i.e., Induction Hypothesis). Now we consider the case of $j = j_0 + 1$. Similar to the proof of Lemma 5, we assume that $\frac{|P_l|}{|P|} \geq \frac{\epsilon}{4j}$ for each $1 \leq l \leq j$. Otherwise, it can be reduced to a problem with smaller j , and solved by the induction hypothesis. Hence, in the following discussion, we assume that $\frac{|P_l|}{|P|} \geq \frac{\epsilon}{4j}$ for each $1 \leq l \leq j$.

First, we know that $o = \sum_{l=1}^j \frac{|P_l|}{|P|} o_l$. Let $o' = \sum_{l=1}^j \frac{|P_l|}{|P|} o'_l$. Then, we have

$$||o - o'|| = \left\| \sum_{l=1}^j \frac{|P_l|}{|P|} o_l - \sum_{l=1}^j \frac{|P_l|}{|P|} o'_l \right\| \leq \sum_{l=1}^j \frac{|P_l|}{|P|} ||o_l - o'_l|| \leq L. \quad (10)$$

Thus, if we can find a grid point τ within a distance to o' no more than $\sqrt{\epsilon}\delta + \epsilon L$ (i.e., $||\tau - o'|| \leq \sqrt{\epsilon}\delta + \epsilon L$), by inequality (10), we will have $||\tau - o|| \leq ||\tau - o'|| + ||o' - o|| \leq \sqrt{\epsilon}\delta + (1 + \epsilon)L$. This means that we only need to find a grid point close enough to o' .

To find such a τ , we first consider the distance from o'_l to o' . For any $1 \leq l \leq j$, we have

$$||o'_l - o'|| \leq ||o'_l - o_l|| + ||o_l - o|| + ||o - o'|| \leq 2\sqrt{\frac{j}{\epsilon}}\delta + 2L, \quad (11)$$

where the first inequality follows from triangle inequality, and the second inequality follows from the facts that $||o'_l - o_l||$ and $||o - o'||$ are both bounded by L , and $||o_l - o|| \leq 2\sqrt{\frac{j}{\epsilon}}\delta$ (by Lemma 4).

This implies that we can use the similar idea in Lemma 5 to construct a ball \mathcal{B} centered at any o'_{l_0} and with radius $r = \max_{1 \leq l \leq j} \{||o'_l - o'_{l_0}||\}$. Note that since $||o'_l - o'_{l_0}|| \leq ||o'_l - o'|| + ||o' - o'_{l_0}|| \leq 4\sqrt{\frac{j}{\epsilon}}\delta + 4L$ (by inequality (11)), the simplex V' is inside \mathcal{B} . Similar to Lemma 5, we can build a grid inside \mathcal{B} with grid length $\frac{\epsilon r}{4j}$ and total grid points $O((8j/\epsilon)^j)$. Clearly in this grid, we can find a grid point τ such that $||\tau - o'|| \leq \frac{\epsilon}{4\sqrt{j}}r \leq \sqrt{\epsilon}\delta + \epsilon L$. Thus, $||\tau - o|| \leq \sqrt{\epsilon}\delta + (1 + \epsilon)L$, and the induction step, as well as the lemma, holds. \square

11 Proof of Claim 1 in Lemma 7

Proof. Since $1 \geq \beta_j \geq \frac{1}{|G|} \geq \frac{1}{kn}$, there is one integer t between 1 and $\log(kn)$, such that $2^{t-1} \leq \frac{1}{\beta_j} \leq 2^t$. Thus $2^{t/2-1}\sqrt{\epsilon}\delta_{opt} \leq \sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt} \leq 2^{t/2}\sqrt{\epsilon}\delta_{opt}$. Together with $\delta \in [\delta_{opt}, (1 + \epsilon)\delta_{opt}]$, we have

$$2^{t/2-1}\sqrt{\epsilon}\frac{\delta}{1 + \epsilon} \leq \sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt} \leq 2^{t/2}\sqrt{\epsilon}\delta.$$

Thus if set $\hat{r}_j = 2^{t/2}\sqrt{\epsilon}\delta$, we have $\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt} \leq \hat{r}_j \leq 2(1+\epsilon)\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}$. Let $x = \frac{j\hat{r}_j}{j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}}$. Then we have $1 \leq x \leq 2(1+\epsilon)$. We build a grid in the interval $[\frac{x}{2(1+\epsilon)}, x]$ with the grid length $\frac{\epsilon}{4(1+\epsilon)}x$, and obtain a grid set (i.e., number set) $\mathcal{N} = \{\frac{1+l\frac{\epsilon}{2}}{2(1+\epsilon)}x \mid 0 \leq l \leq 4 + \frac{2}{\epsilon}\}$. We prove that there must exist one number in \mathcal{N} and is between 1 and $1 + \epsilon/2$. First, we know that $\frac{x}{2(1+\epsilon)} \leq 1 \leq x$. If $x \leq 1 + \epsilon/2$, we find the the desired number x in \mathcal{N} . Otherwise, the whole interval $[1, 1 + \epsilon/2]$ is inside $[\frac{x}{2(1+\epsilon)}, x]$. Since the grid has grid length $\frac{\epsilon}{4(1+\epsilon)}x \leq \frac{\epsilon}{4(1+\epsilon)}2(1+\epsilon) = \epsilon/2$, there must exist one grid point locating inside $[1, 1 + \epsilon/2]$. Thus, the desired number exists in \mathcal{N} .

Let $\mathcal{R}_j = \{\frac{1+l\frac{\epsilon}{2}}{2(1+\epsilon)}j\hat{r}_j \mid 0 \leq l \leq 4 + \frac{2}{\epsilon}\}$. From the above analysis, we know that there exists one value $r_j \in \mathcal{R}_j$ such that

$$j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt} \leq r_j \leq (1 + \frac{\epsilon}{2})j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}.$$

Note that $\mathcal{R}_j \subset \mathcal{R}$, where $\mathcal{R} = \cup_{t=0}^{\log(kn)} \{\frac{1+l\frac{\epsilon}{2}}{2(1+\epsilon)}j2^{t/2}\sqrt{\epsilon}\delta \mid 0 \leq l \leq 4 + \frac{2}{\epsilon}\}$. Thus, the Claim is proved. \square

12 Proof of Claim 2 in Lemma 7

Proof. First, for each $1 \leq l \leq j-1$, we have $|Opt_l \setminus (\cup_{w=1}^{j-1} B_{j,w})| \leq |Opt_l \setminus B_{j,l}|$. Secondly, by Markov inequality, we have

$$|Opt_l \setminus B_{j,l}| \leq \frac{\delta_l^2}{(r_j - \|p_{v_l} - m_l\|)^2} |Opt_l|.$$

Note that $\delta_{opt}^2 = \sum_{j=1}^k \beta_j \delta_j^2$, and $\beta_j \leq \beta_l$ (by $l < j$). Thus, we have $\delta_l \leq \sqrt{\frac{1}{\beta_l}}\delta_{opt} \leq \sqrt{\frac{1}{\beta_j}}\delta_{opt}$. Together with $j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt} \leq r_j$ and $\|p_{v_l} - m_l\| \leq \epsilon\delta_l + (1+\epsilon)l\sqrt{\frac{\epsilon}{\beta_l}}\delta_{opt}$ (by induction hypothesis), we have

$$\begin{aligned} r_j - \|p_{v_l} - m_l\| &\geq j\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt} - (\epsilon\delta_l + (1+\epsilon)(j-1)\sqrt{\frac{\epsilon}{\beta_l}}\delta_{opt}) \\ &= (1 - (j-1)\epsilon)\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt} - \epsilon\delta_l \\ &\geq (1 - (j-1)\epsilon - \sqrt{\epsilon})\sqrt{\frac{\epsilon}{\beta_j}}\delta_{opt}, \end{aligned}$$

where the last inequality follows from $\delta_l \leq \sqrt{\frac{1}{\beta_l}}\delta_{opt} \leq \sqrt{\frac{1}{\beta_j}}\delta_{opt}$. Thus, we have

$$\begin{aligned} |Opt_l \setminus B_{j,l}| &\leq \frac{\delta_l^2}{(1 - (j-1)\epsilon - \sqrt{\epsilon})^2 \frac{\epsilon}{\beta_j} \delta_{opt}^2} |Opt_l| \\ &\leq \frac{\delta_l^2}{(1 - (j-1)\epsilon - \sqrt{\epsilon})^2 \frac{\epsilon}{\beta_j} \beta_l \delta_l^2} |Opt_l| \\ &= \frac{\beta_j}{(1 - (j-1)\epsilon - \sqrt{\epsilon})^2 \epsilon \beta_l} |Opt_l| \\ &= \frac{\beta_j |\mathcal{G}|}{(1 - (j-1)\epsilon - \sqrt{\epsilon})^2 \epsilon} \leq \frac{\beta_j |\mathcal{G}|}{(1 - j\sqrt{\epsilon})^2 \epsilon}, \end{aligned}$$

where the second inequality follows from the fact that $\beta_l \delta_l^2 \leq \delta_{opt}^2$, and the fourth equation follows from that $\frac{|Opt_l|}{\beta_l} = |\mathcal{G}|$. Note that we can assume ϵ is small enough such that $\epsilon \leq \frac{1}{4j^2}$, which implies

that $\frac{\beta_j |\mathcal{G}|}{(1-j\sqrt{\epsilon})^2 \epsilon} \leq \frac{4\beta_j |\mathcal{G}|}{\epsilon}$. Otherwise, we can just replace ϵ by $\frac{\epsilon}{4j^2}$ as the input at the beginning of the algorithm. Thus, in total, we have

$$|Opt_l \setminus B_{j,l}| \leq \frac{4\beta_j |\mathcal{G}|}{\epsilon}.$$

Thus the Claim is proved. \square

13 Proof of Claim 3 in Lemma 7

Proof. First, we have

$$\frac{|P_1|}{|P_1 \cup P_j|} \geq \frac{\frac{1}{j-1}(1 - \frac{\epsilon^3}{j})}{\frac{1}{j-1}(1 - \frac{\epsilon^3}{j}) + \frac{\epsilon^3}{j}} > \frac{1 - \epsilon^3}{1 + \epsilon^3}.$$

Let a^2 denote the variance of $P_1 \cup P_j$. By Lemma 4, we know that $\|\tau_1 - \tau'\| \leq \sqrt{\frac{2\epsilon^3}{1-\epsilon^3}} a$. Meanwhile, since $\frac{|P_1 \cup P_j|}{|Opt_j|} \geq \frac{|P_1|}{|Opt_j|} \geq \frac{\frac{1}{j-1}(\beta_j - \epsilon^3 \frac{\beta_j}{j})|\mathcal{G}|}{\beta_j |\mathcal{G}|} = \frac{1 - \frac{\epsilon^3}{j}}{j-1}$, we have $a^2 \leq \frac{|Opt_j|}{|P_1 \cup P_j|} \delta_j^2 \leq \frac{j-1}{1 - \frac{\epsilon^3}{j}} \delta_j^2$. Then we have

$$\begin{aligned} \|\tau_1 - \tau'\| &\leq \sqrt{\frac{2\epsilon^3}{1-\epsilon^3}} a \leq \sqrt{\frac{2\epsilon^3}{1-\epsilon^3}} \sqrt{\frac{j-1}{1 - \frac{\epsilon^3}{j}}} \delta_j \\ &\leq \sqrt{\frac{2j\epsilon^3}{(1-\epsilon^3)(1 - \frac{\epsilon^3}{j})\beta_j}} \delta_{opt} \\ &\leq \sqrt{\frac{2j\epsilon^3}{(1-\epsilon^3)(1 - \epsilon^3)\beta_j}} \delta_{opt} = \frac{\sqrt{2}\epsilon}{1 - \epsilon^3} \sqrt{\frac{j\epsilon}{\beta_j}} \delta_{opt}, \end{aligned}$$

where the third inequality follows from $\delta_j \leq \sqrt{\frac{1}{\beta_j}} \delta_{opt}$. Thus, the claim is true. \square

14 Chromatic k -Medians Clustering

In this section, we extend our ideas for k -CMeans to the Chromatic k -Medians Clustering problem (k -CMedians). Similar to k -CMeans, we first show its relationship with k -medians (in Section 14.1), and then present a $(5 + \epsilon)$ -approximation algorithm (in Section 14.2). Due to the lack of a similar Simplex Lemma for k -CMedians, we achieve a constant approximation, instead of a PTAS.

Definition 3 (Chromatic k -Median Clustering (k -CMedians)). Let $\mathcal{G} = \{G_1, \dots, G_n\}$ be a set of n point-sets with each $G_i = \{p_1^i, \dots, p_{k_i}^i\}$ consisting of $k_i \leq k$ points in \mathbb{R}^d space. The chromatic k -median clustering (or k -CMedians) of \mathcal{G} is to find k points $\{m_1, \dots, m_k\}$ in \mathbb{R}^d space and a chromatic partition U_1, \dots, U_k of \mathcal{G} such that $\frac{1}{n} \sum_j \sum_{q \in U_j} \|q - m_j\|$ is minimized.

14.1 Constant Approximation from k -Medians

Given a set of points in \mathbb{R}^d , the optimal median point is also called *Fermat Weber point* in geometry. Its main difference with mean point is that no explicit formula exists for computing the optimal median point, while the mean point is simply the average of the given points. Consequently, median point is often approximated using some iterative procedure, such as *Weiszfeld's algorithm*. Thus in the following discussion, we only assume the availability of a $(1 + \epsilon)$ -approximation of the median point.

Lemma 9. Let P be a set of points in \mathbb{R}^d space, and P_1 be a subset of P containing a fraction of $\alpha \leq 1$ points of P . Let m_{opt} and m be the optimal and $(1 + \epsilon)$ -approximate median point of P respectively, and m_1 be the optimal median of P_1 . Then $\|m_1 - m\| \leq \frac{2+\epsilon}{\alpha}\mu$, where $\mu = \frac{1}{|P|} \sum_{p \in P} \|p - m_{opt}\|$.

Proof. Let $\mu_1 = \frac{1}{|P_1|} \sum_{p \in P_1} \|p - m_1\|$. Since $P_1 \subseteq P$, it is easy to know that $\sum_{p \in P} \|p - m\| \geq \sum_{p \in P_1} \|p - m\|$, which implies that $(1 + \epsilon)\mu \geq \frac{1}{|P|} \sum_{p \in P_1} \|p - m\| = \alpha \frac{1}{|P_1|} \sum_{p \in P_1} \|p - m\|$. By triangle inequality, we also have $\|p - m\| \geq \|m - m_1\| - \|p - m_1\|$. Thus,

$$(1 + \epsilon)\mu \geq \alpha(\|m - m_1\| - \mu_1). \quad (12)$$

Since m_1 is the optimal median of P_1 , we have $\mu_1 = \frac{1}{|P_1|} \sum_{p \in P_1} \|p - m_1\| \leq \frac{1}{|P_1|} \sum_{p \in P_1} \|p - m_{opt}\| \leq \frac{1}{|P_1|} \sum_{p \in P} \|p - m_{opt}\| = \frac{1}{\alpha}\mu$. Plugging this into inequality (12), we have $\|m - m_1\| \leq \frac{2+\epsilon}{\alpha}\mu$. \square

Theorem 6. Let $\mathcal{G} = \{G_1, \dots, G_n\}$ be an instance of k -CMedians, and \mathcal{C} be the $k(1 + \epsilon)$ -approximate median points of the k clusters generated by a c -approximation k -medians algorithm on the points $\cup_{i=1}^n G_i$. Then, $[\mathcal{C}]^k$ contains at least one k -tuple whose elements are the k median points of a $((2 + \epsilon)ck^2 + (2 + \epsilon)k + 1)$ -approximation of k -CMedians on \mathcal{G} , where $[\mathcal{C}]^k = \underbrace{\mathcal{C} \times \dots \times \mathcal{C}}_k$.

Proof. Let $\{c_1, \dots, c_k\}$ be the set of k approximate median points in \mathcal{C} , and $\{S_1, \dots, S_k\}$ be the k clusters returned by the c -approximation k -medians algorithm. Thus, c_j is the $(1 + \epsilon)$ -approximate median point of S_j for $1 \leq j \leq k$. Let $\mathcal{OPT} = \{Opt_1, \dots, Opt_k\}$ be the **unknown** optimal solution for k -CMedians on \mathcal{G} , and m_j be the optimal median point of Opt_j for $1 \leq j \leq k$. Denote the set $Opt_i \cap S_j$ as Γ_j^i , and its optimal median point as $\tau_{j_i}^i$ for $1 \leq i, j \leq k$.

Since $\cup_{j=1}^k \Gamma_j^i = Opt_i$, there must exist some index $1 \leq j_i \leq k$ such that $|\Gamma_{j_i}^i| \geq \frac{1}{k}|Opt_i|$. Fixing j_i , we have the following about $\sum_{p \in Opt_i} \|p - c_{j_i}\|$.

$$\begin{aligned} \sum_{p \in Opt_i} \|p - c_{j_i}\| &= \sum_{p \in Opt_i} \|p - m_i + m_i - c_{j_i}\| \\ &\leq \sum_{p \in Opt_i} (\|p - m_i\| + \|m_i - c_{j_i}\|) \\ &= \sum_{p \in Opt_i} \|p - m_i\| + |Opt_i| \times \|m_i - c_{j_i}\| \\ &= \sum_{p \in Opt_i} \|p - m_i\| + |Opt_i| \times \|m_i - \tau_{j_i}^i + \tau_{j_i}^i - c_{j_i}\| \\ &\leq \sum_{p \in Opt_i} \|p - m_i\| + |Opt_i| \times (\|m_i - \tau_{j_i}^i\| + \|\tau_{j_i}^i - c_{j_i}\|) \end{aligned}$$

By Lemma 9, we have

$$\begin{aligned} \|\tau_{j_i}^i - m_i\| &\leq \frac{2 + \epsilon}{\frac{1}{k}} \left(\frac{1}{|Opt_i|} \sum_{p \in Opt_i} \|p - m_i\| \right); \\ \|\tau_{j_i}^i - c_{j_i}\| &\leq \frac{2 + \epsilon}{\frac{|\Gamma_{j_i}^i|}{|S_{j_i}|}} \left(\frac{1}{|S_{j_i}|} \sum_{p \in S_{j_i}} \|p - c_{j_i}\| \right). \end{aligned}$$

From the above inequalities, we have

$$\begin{aligned}
\sum_{p \in Opt_i} \|p - c_{j_i}\| &\leq \sum_{p \in Opt_i} \|p - m_i\| + |Opt_i| \times (\|m_i - \tau_{j_i}^i\| + \|\tau_{j_i}^i - c_{j_i}\|) \\
&\leq \sum_{p \in Opt_i} \|p - m_i\|^2 + |Opt_i| \left(\frac{2+\epsilon}{\frac{1}{k}} \left(\frac{1}{|Opt_i|} \sum_{p \in Opt_i} \|p - m_i\| \right) + \frac{2+\epsilon}{\frac{|\Gamma_{j_i}^i|}{|S_{j_i}|}} \left(\frac{1}{|S_{j_i}|} \sum_{p \in S_{j_i}} \|p - c_{j_i}\| \right) \right) \\
&= ((2+\epsilon)k+1) \sum_{p \in Opt_i} \|p - m_i\| + (2+\epsilon) \frac{|Opt_i|}{|\Gamma_{j_i}^i|} \times \sum_{p \in S_{j_i}} \|p - c_{j_i}\|.
\end{aligned}$$

Since $|\Gamma_{j_i}^i| \geq \frac{1}{k}|Opt_i|$, we have $\frac{|Opt_i|}{|\Gamma_{j_i}^i|} \leq k$. Thus,

$$\sum_{p \in Opt_i} \|p - c_{j_i}\|^2 \leq ((2+\epsilon)k+1) \sum_{p \in Opt_i} \|p - m_i\| + (2+\epsilon)k \sum_{p \in S_{j_i}} \|p - c_{j_i}\|.$$

Summing both sides of the above inequality over i , we have

$$\begin{aligned}
\sum_{i=1}^k \sum_{p \in Opt_i} \|p - c_{j_i}\|^2 &\leq ((2+\epsilon)k+1) \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\| + (2+\epsilon)k \sum_{i=1}^k \sum_{p \in S_{j_i}} \|p - c_{j_i}\| \\
&\leq ((2+\epsilon)k+1) \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\| + (2+\epsilon)k^2 \sum_{j=1}^k \sum_{p \in S_j} \|p - c_j\|. \quad (13)
\end{aligned}$$

It is easy to know that the optimal objective value of k -medians is no larger than that of k -CMedians on the same set of points in \mathcal{G} . Thus, $\sum_{j=1}^k \sum_{p \in S_j} \|p - c_j\| \leq c \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\|$. Plugging this inequality into (13), we have

$$\sum_{i=1}^k \sum_{p \in Opt_i} \|p - c_{j_i}\| \leq ((2+\epsilon)ck^2 + (2+\epsilon)k+1) \sum_{i=1}^k \sum_{p \in Opt_i} \|p - m_i\|.$$

The above inequality means that if we take the k -tuple $(c_{j_1}, \dots, c_{j_k})$ as the k approximate median points for k -CMedians, we have a $((2+\epsilon)ck^2 + (2+\epsilon)k+1)$ -approximation solution for k -CMedians. Thus, the theorem is proved. \square

14.2 Peeling Algorithm for k -CMedians

The following lemma is a key to the peeling algorithm for k -CMedians (i.e., play a similar role as Lemma 5 for k -CMeans).

Lemma 10. *Let P to be a set of points in \mathbb{R}^d with a partition $P = \cup_{l=1}^j P_l$, o be its optimal median point, and o_l be the optimal median point of P_l for $1 \leq l \leq j$. Let $\mu = \frac{1}{|P|} \sum_{p \in P} \|p - o\|$. Then, there exists some i_0 such that $\|o - o_{i_0}\| \leq 4\mu$.*

Proof. Since $\mu = \frac{\sum_{p \in P} \|p - o\|}{|P|} = \sum_{i=1}^l \left(\frac{|P_i|}{|P|} \frac{\sum_{p \in P_i} \|p - o\|}{|P_i|} \right)$, there must exist some index i_0 such that $\frac{\sum_{p \in P_{i_0}} \|p - o\|}{|P_{i_0}|} \leq \mu$. By Markov inequality, we know that there exists one subset U of P_{i_0} such that $|U| > |P_{i_0}|/2$ and $\|p - o\| \leq 2\mu$ for any $p \in U$.

Since o_{i_0} is the optimal median point of P_{i_0} , $\frac{\sum_{p \in P_{i_0}} \|p - o_{i_0}\|}{|P_{i_0}|} \leq \frac{\sum_{p \in P_{i_0}} \|p - o\|}{|P_{i_0}|} \leq \mu$. Similarly, by Markov inequality, we know that there exists one subset V of P_{i_0} such that $|V| > |P_{i_0}|/2$ and $\|p - o_{i_0}\| \leq 2\mu$ for any $p \in V$.

From the inequalities of $|U| > |P_{i_0}|/2$ and $|V| > |P_{i_0}|/2$ and the fact that $U \cap V \neq \emptyset$, we know that there exists one point $p_0 \in U \cap V$ such that $\|p_0 - o\| \leq 2\mu$ and $\|p_0 - o_{i_0}\| \leq 2\mu$. Thus $\|o_{i_0} - o\| \leq \|o_{i_0} - p_0\| + \|p_0 - o\| \leq 4\mu$. \square

Before presenting our peeling algorithm, we still need the following lemma proved by Badoiu *et al.* in [6] for finding an approximate solution for 1-median.

Theorem 7 ([6]). *Let P be a normalized set of n points in \mathbb{R}^d space, $1 > \epsilon > 0$, and R be a random sample of $O(1/\epsilon^3 \log 1/\epsilon)$ points from P . Then one can compute, in $O(d2^{O(1/\epsilon^4)} \log n)$ time, a point-set $S(P, R)$ of cardinality $O(2^{O(1/\epsilon^4)} \log n)$, such that with constant probability (over the choices of R), there is a point $q \in S(P, R)$ such that $\text{cost}(q, P) \leq (1 + \epsilon)\text{med}_{\text{opt}}(P, 1)$.*

Algorithm k -CMedians

Input: $\mathcal{G} = \{G_1, \dots, G_n\}$, $k \geq 2$ and an small constant $\epsilon > 0$.

Output: a $(5 + \epsilon)$ -approximation solution for k -CMedians on \mathcal{G} .

1. Run the $(1 + \epsilon)$ -approximation k -medians algorithm from [15] on \mathcal{G} , and let Ω be the obtained objective value.
2. For $i = 1$ to $\frac{4k}{\epsilon}$ do
 - (a) Set $\delta = \frac{\Omega}{4k} + i \frac{\epsilon}{4k} \Omega$, and run Algorithm Sphere-Peeling-Tree-2.
 - (b) Let \mathcal{T}_i be the returned tree.
3. For each path of every \mathcal{T}_i , use bipartite matching procedure to compute the objective value of k -CMeans on \mathcal{G} . Output the k points from the path with smallest objective value.

Algorithm Sphere-Peeling-Tree-2

Input: \mathcal{G} , $k \geq 2$, $\epsilon, \delta > 0$.

Output: A tree \mathcal{T} of height k with each node v associated with a point $p_v \in \mathbb{R}^d$.

1. Initialize \mathcal{T} with a single root node v associating with no point.
2. Recursively grow each node v in the following way
 - (a) If the height of v is already k , then it is a leaf node.
 - (b) Otherwise, let j be the height of v . Build the set of radius candidates $\mathcal{R} = \cup_{t=0}^{\log(kn)} \{ \frac{1+t\epsilon}{2(1+\epsilon)} j 2^{t/2} \sqrt{\epsilon} \delta \mid 0 \leq l \leq 4 + \frac{2}{\epsilon} \}$. For each radius candidate $r \in \mathcal{R}$ do
 - i. Let j be the height of v , and $\{p_{v_1}, \dots, p_{v_j}\}$ be the j points associated with nodes on the root-to- v path (including p_v).
 - ii. For each p_{v_l} , $1 \leq l \leq j$, construct a ball $B_{j,l}$ centered at p_{v_l} and with radius r .
 - iii. Take a random sample from $\mathcal{G} \setminus \cup_{l=1}^j B_{j,l}$ with size $m = \frac{8k^3}{\epsilon^9} \ln \frac{k^2}{\epsilon^6}$. Compute the approximate median points of all subsets of the sample (by Theorem 7), and denote the set of the approximate median points as Π . Clearly, $|\Pi| = 2^{m+O(1/\epsilon^4)} \log n$.
 - iv. For each point p in Π , add one child to v , and associate it with p ; add another j children, with each one associating with a different point in $\{p_{v_1}, \dots, p_{v_j}\}$.

We can use a similar approach as in Section 5.3 to analyze the correctness of Algorithm k -CMedians.

Let $\text{OPT} = \{\text{Opt}_1, \dots, \text{Opt}_k\}$ be the optimal solution of k -CMedians on \mathcal{G} . Without loss of generality, we assume that $|\text{Opt}_1| \geq |\text{Opt}_2| \geq \dots \geq |\text{Opt}_k|$. For each Opt_j , $1 \leq j \leq k$, let m_j be its median point, β_j be its fraction in \mathcal{G} (i.e., $|\text{Opt}_j|/|\cup_{i=1}^n G_i|$), and $\mu_j = \frac{1}{|\text{Opt}_j|} \sum_{p \in \text{Opt}_j} \|p - m_j\|$. Thus, $\beta_1 \geq \dots \geq \beta_k$ and $\sum_{j=1}^k \beta_j = 1$. Also, let $\mu_{\text{opt}} = \sum_{j=1}^k \beta_j \mu_j$.

Lemma 11. *Among all the trees generated in Algorithm k -CMedians, there exists one tree \mathcal{T}_i , which has a root-to-leaf path with each node v_j at level j , $1 \leq j \leq k$, on the path associating a point p_{v_j} and satisfying the inequality*

$$\|p_{v_j} - m_j\| \leq 4\mu_j + (1 + \epsilon)j \frac{\epsilon}{\beta_j} \mu_{opt}.$$

Lemma 12. *If Lemma 11 is true, Algorithm k -CMedians yields a $(5 + O(k^2)\epsilon)$ -approximation solution for k -CMedians.*

Proof. We first assume that Lemma 11 is true. Then for each $1 \leq j \leq k$, we have

$$\begin{aligned} \sum_{p \in Opt_j} \|p - p_{v_j}\| &\leq \sum_{p \in Opt_j} \|p - m_j\| + |Opt_j| \times \|m_j - p_{v_j}\| \\ &\leq \sum_{p \in Opt_j} \|p - m_j\| + |Opt_j| \times (4\mu_j + (1 + \epsilon)j \frac{\epsilon}{\beta_j} \mu_{opt}) \\ &= 5|Opt_j|\mu_j + (1 + \epsilon)j\epsilon|\mathcal{G}|\mu_{opt} \end{aligned} \tag{14}$$

Summing the both sides of (14) over j , we have

$$\begin{aligned} \sum_{j=1}^k \sum_{p \in Opt_j} \|p - p_{v_j}\|^2 &\leq \sum_{j=1}^k (5|Opt_j|\mu_j + (1 + \epsilon)j\epsilon|\mathcal{G}|\mu_{opt}) \\ &\leq 5 \sum_{j=1}^k |Opt_j|\mu_j + (1 + \epsilon)k^2\epsilon|\mathcal{G}|\mu_{opt} \\ &= (5 + O(k^2)\epsilon)|\mathcal{G}|\mu_{opt}. \end{aligned} \tag{15}$$

In the above, the last equation follows from the fact that $\sum_{j=1}^k |Opt_j|\mu_j = |\mathcal{G}|\mu_{opt}$. By (15), we know that $\{p_{v_1}, \dots, p_{v_k}\}$ induces a $(5 + O(k^2)\epsilon)$ -approximation solution for k -CMedians. \square

By a similar argument given in the proof of Lemma 7, we can show the correctness of Lemma 11. Thus, we have the following theorem.

Theorem 8. *With constant probability, Algorithm k -CMedians yields a $(5 + \epsilon)$ -approximation for k -CMedians in $O(2^{\text{poly}(\frac{k}{\epsilon})}n(\log n)^{2k}d)$ time.*